ISSN: 2579-5643
ISBN: 978-602-74355-1-3

Klasifikasi Abstrak Tugas Akhir Mahasiswa DIII Politeknik Harapan Bersama Tegal

Yerry Febrian Sabanise[1*)]
[1]Jurusan Teknik Komputer, Politeknik Harapan Bersama Tegal
email: [1]yryfebrian@gmail.com,

Abstrak **J Abstraksi tugas akhir mahasiswa merupakan initisari dari suatu penelitan yang dilakukan oleh mahasiswa. Berbagai tema diangkat dalam tugas akhir ini. Tetapi dari tema tema tersebut, untuk mengklasifikasi abstrak tugas akhir mahasiswa masih sulit dilakukan dilihat dari akurasi penelitian yang telah dilakukan belum mencapai 90%. Oleh karena itu, penelitian ini dilakukan untuk mengklasifikasi abstrak tugas akhir untuk mendapatkan memudahkan dalam mencari tugas akhir. Dan juga untuk dapat menentukan atribut terbaik dari hasil *text prosesing* dengan klasifikasi menggunakan naive bayes. permasalahannya adalah tidak adanya pedoman baku dalam menentukan parameter yang akan digunakan pada metode ini sehingga yang dipakai adalah metode eksperimen. Untuk itu diperlukan metode yang dapat menyelesaikan permasalahan tersebut, sehingga parameter yang didapatkan dapat menjadi lebih optimal. Solusi yang dapat diterapkan adalah dengan menerapkan Algoritma genetika (GA) pada *Naïve Bayes,* untuk dapat menentukan atribut terbaik. Hasil yang didapatkan adalah ternyata penerapan teknik optimasi dengan Algoritma Genetika dapat mempermudah dalam mencari nilai parameter secara optimal dan dapat meningkatkan nilai akurasi pada algoritma *Naïve bayes,* dengan demikian model yang didapatkan dapat digunakan bagi para pencari referensi tugas akhir untuk mencari referensi tugas akhir yang tepat berdasarkan kata dari atribut yang terbaik.**

Kata Kunci : *text prosesing*, *Naive bayes*, Abstrak, algoritma genetika(GA)

I. PENDAHULUAN

Ada anggapan sebagian mahasiswa, bahwa menyusun Tugas Akhir dengan bahasa yang baik dan benar itu rumit dan menyusahkan. Sebagian mereka itu mengeluh setelah diberi tugas menyusun makalah atau tugas akhir oleh dosen pembimbing atau lembaga pendidikan tingginya. Mereka seakan-akan "menyerah" sebelum "bertempur"[1].

Tugas Akhir (TA) adalah hasil tertulis dari pelaksanaan suatu penelitian, yang dibuat untuk pemecahan masalah tertentu dengan menggunkan kaidah-kaidah yang berlaku dalam bidang ilmu tersebut. [2]

Klasifikasi dan prediksi adalah dua bentuk analisis data yang dapat digunakan untuk mengekstrak model dari data yang berisi kelas-kelas atau untuk memprediksi trend data yang akan datang. Klasifikasi memprediksi data dalam bentuk kategori, sedangkan prediksi memodelkan fungsi-fungsi dari nilai yang lanjut [3].

Klasifikasi Bayes juga dikenal dengan Naive Bayes, memiliki kemampuan sebanding dengan pohon keputusan. Naive bayes merupakan salah satu penerapan teorema bayesian. Naive bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output [4]. Naive Bayes dapat menggunakan penduga kernel kepadatan, yang meingkatkan kinerja jika asumsi normalitas sangat tidak benar, tetapi juga dapat menangani atribut numeric menggunakan dikstritisasi diawasi [5].

Klasifikasi abstrak tugas akhir menerapkan seleksi atribut algoritma genetika pada metode klasifikasi *naive bayes* yang di targetkan pada empat kategori bidang ilmu yaitu bidang ilmu komputer, bidang ilmu farmasi, bidang ilmu akuntansi, bidang ilmu kebidanan.

Tujuan dari penelitian ini adalah mengetahui tingkat akurasi seleksi atribut dalam klasifikasi abstrak tugas akhir dengan menggunakan algoritma *naive bayes* dan algoritma genetika pada empat kategori bidang ilmu yaitu bidang ilmu komputer, bidang ilmu farmasi, bidang ilmu akuntansi, bidang ilmu kebidanan.

Untuk mendapatkan nilai akurasi yang baik pada Naive Bayes, GA yang merupakan salah satu algoritma terbaik untuk optimasi, digunakan sebagai model untuk mendapatkan nilai akurasi Naive bayes yang terbaik. Dengan diterapkannya algoritma genetika pada Naive bayes diharapkan dapat mempercepat proses pencarian dalam mendapatkan nilai akurasi yang sesuai dan optimal pada Naive Bayes sehingga dapat meningkatkan tingkat akurasi Text mining dalam mengklasifikasi abstrak tugas akhir.

II. TINJAUAN STUDI

A. Penelitian Terkait

Penelitian dilakukan oleh vivek Narayan, Ishan Arora, Arjun Bhatia tahun 2013 [6], setelah menjelajahi metode yang berbeda untuk meningkatkan akurasi dari Bayes classifier naif untuk analisis sentimen. Kami mengamati bahwa kombinasi metode seperti penanganan efektif negasi, kata n-gram dan seleksi fitur dengan hasil informasi timbal balik dalam peningkatan yang signifikan dalam akurasi. Ini berarti bahwa sangat akurat dan cepat sentimen classifier dapat dibangun dengan menggunakan model Naive Bayes sederhana yang memiliki pelatihan linear dan waktu pengujian hubungan complexi-. Dapat mencapai akurasi 88,80% pada populer IMDB review film dataset.

Penelitian dilakukan oleh Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari, and Jordan Pascual (2014) [7], pada penelitian kali ini peneliti mencoba untuk mencari dari beberapa information retrival dan machine learning yang ada lebih baik mana dalam mengklasifikasi tekt, metode yang

*) penulis korespondensi

ISSN: 2579-9045
ISBN: 978-602-74355-1-3

digunakan dalam penelitian ini adlah KNN dengan naive Bayes dengan mnggunakan data Reuters-21578 dataset yang tersedia dengan 21.578 dokumen. Kesimpulan yang diambil dari membanding kan 2 metode yakni metode KNN dan Naive bayes, metode yang mempunyai akurasi yang lebih besar adalah KNN dengan 90% pada kategory masyarakat.

Penelitian dilakukan oleh Ms S Vijayarani, Ms M Muthulakshmi 2013 [8], menganalisis kinerja pengklasifikasi Bayesian dan lazy classifiers file yang disimpan dalam hard disk komputer. Ada dua algoritma di classifier Bayesian yaitu BayesNet, dan Naïve Bayes. Dalam lazy classifier memiliki tiga algoritma yaitu IBL, IBK dan Kstar. Kinerja pengklasifikasi Bayesian dan lazy classifier dianalisis dengan menerapkan berbagai faktor kinerja. Dari hasil percobaan, teramati bahwa lazy classifier malas lebih efisien daripada classifier Bayesian. Dengan hasil RMSE IBL 6.97, IBK 5,95, Kstar 14,9.

B. Landasan Teori

Landasan teori berisi mengenai klasifikasi dan tahap text mining untuk menemukan atau menggali informasi dari kumpulan dokumen teks yang besar. Dalam *text mining* memiliki lima tahapan dalam pemrosesan data teks tetapi pada tugas akhir ini akan hanya menggunakan 4 tahap yang antara lain [9] :

1) Tahap Tokenizing adalah tahap pemotongan *string input* berdasarkan tiap kata yang menyusunnya.
2) Tahap Filtering adalah tahap mengambil kata-kata penting dari hasil token. Tahap ini biasanya juga disebut tahap *stopword removal*.
3) Tahap Stemming adalah tahap mentransformasi kata-kata hasil *filtering* ke kata-kata akarnya (*root word*) atau kata dasar dengan menggunakan aturan-aturan tertentu.
4) Tahap Analyzing adalah Tahap ini merupakan tahap penentuan seberapa jauh keterkaitan antar kata-kata pada dokumen/inputan yang ada. Pada tahap *anlyzing* akan digunakan rumus TF-IDF untuk mengambil sebuah informasi dari sebuah dokumen. Kata-kata yang umum dalam sebuah dokumen cenderung memiliki nilai tinggi dalam perhitungan TF-IDF.

$$tfidf(w) = tf \times log \frac{n}{df(w)} \quad \text{..... (1)}$$

Teknik Naive Bayes adalah salah satu bentuk sederhana dari Bayesian yang jaringan untuk klasifikasi. Untuk menghasilkan nilai probabitas pada sebuah sampel diberikan sebuah teorema bayesian :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{..... (2)}$$

Sedangkan perhitungan probabilitas setiap atribut menggunakan:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{..... (3)}$$

C. Kerangka Pemikiran

Kerangka Pemikiran berisi mengenai masih banyaknya kesalahan-kesalahan dalam kesesuaian pembuatan abstrak tugas akhir yang membuat kurang kesesuaian abstrak dengan 4 kategori bidang yang ada. Penelitian-penelitian yang pernah melakukan pengkategorian atau pengklasifikasian baik dalam bentuk dokumen maupun bentuk teks mengatakan bahwa hasil akurasi rata-rata di atas 70% dengan menggunakan algoritma-algoritma klasifikasi *machine learning* salah satunya adalah algoritma Naïve Bayes. Oleh karena itu pada penelitian ini akan menggunakan algoritma Naïve Bayes sebagai algoritma klasifikasi dan algoritma genika untuk seleksi atributnya. Penggunaan algoritma seleksi atribut *algoritma genetic* bertujuan untuk meningkatkan akurasi hasil klasifikasi dengan metode Naïve Bayes menggunakan data abstrak tugas akhir yang akan diklasifikasi ke dalam empat kategori yaitu, bidang komputer, bidang kebidanan, bidang farmasi, bidang akuntansi.

III. METODE PENELITIAN

A. Pengumpulan Data

Pada bagian metode pengumpulan data dijelaskan Dataset yang digunakan dalam penelitian ini berupa data abstrak tugas akhir yang terdiri dari empat kategori yaitu bidang ilmu komputer, bidang ilmu farmasi, bidang ilmu akuntansi, bidang ilmu kebidanan [10]. Data tersebut diambil dari perpustakaan Politeknik Harapan Bersama Tegal. Jumlah dataset yang digunakan sebanyak 200 abstrak yang dibagi masing masing 50 data abstrak. Dataset diklasifikasi berdasarkan isi abstrak setiap kategori .

B. Pengolahan Data Awal

Pengolahan data awal (*preprocessing)* merupakan tahap untuk mempersiapkan data yang telah diperoleh dari tahap pengumpulan data yang akan digunakan pada tahap selanjutnya. Proses selanjutnya yaitu:

1) Tahap Tokenize (Term) adalah tahap pemotongan teks menjadi beberapa bagian. Yang bertujuan memisahkan kata per kata dari kalimat abstrak. Term dapat berupa kata atau frasa di dalam dokumen. Namun, kata-kata yang tidak memberikan perbedaan seperti ini, itu, saya, kamu, serta tanda-tanda baca dihilangkan atau dianggap bukan *term.* Hal ini bertujuan untuk mendapatkan hanya kata – kata tertentu saja yang nantinya didapat dan berkontribusi sebagai ciri – ciri dari masing-masing jenis bidang ilmu.
2) Stopword adalah Tahap yang dilakukan dalam *stopword* adalah membuang kata – kata yang tidak penting caranya dengan mencocokan dengan stopword atau kata dasar yang ada pada kata bahasa indonesia. Proses stopwords ini membuat kamus *stopwords* sendiri sebanyak 31949 kata yang tentunya masih sedikit jika dibandingkan dengan jumlah kata pada bahasa Indonesia.
3) Stemming dilakukan dalam penelitian ini untuk mendapatkan kata dasar, dikarenakan ditiap dokumen tugas akhir banyak terdapat kata yang memiliki banyak imbuhan. Contohnya kata "penolong", "ditolong", "pertolongan" memiliki kata dasar yang sama yaitu

ISSN: 2579-9045
ISBN: 978-602-74355-1-3

"tolong". Dalam proses *Stemming* inilah nantinya tiap kata-yang memiliki kata dasar yang sama akan dihilangkan imbuhannya sehingga didapat hanya kata dasar saja. Proses *Stemming* ini dilakukan dengan cara manual yaitu dengan membuat kamus *Stemming* sendiri sebanyak 7185 kata yang tentunya masih sedikit jika dibandingkan dengan jumlah kata pada bahasa Indonesia.

4) Term Weighting, pada tahap ini peneliti menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF adalah metode pembobotan yang mengaitkan antara term frequency (TF) dan inverse document frequency (IDF) [9]. Dalam penelitian ini sebuah kata perlu di beri bobot, makin sering suatu kata muncul pada suatu dokumen, maka diduga kata tersebut semakin penting dalam dokumen.

C. Metode yang Diusulkan

Metode yang diusulkan adalah penggunaan naive bayes untuk mengklasifikasi abstrak tugas akhir. Untuk meningkatkan akurasi klasifikasi abstrak tugas akhir menerapkan algoritma genetika (GA) untuk seleksi atributnya.

D. Evaluasi dan Hasil Validasi

Hasil dari penelitian klasifikasi ini untuk mendapatkan akurasi klasifikasi terbaik. Untuk menentukan nilai tingkat *Accuracy, precesion, recall* pengujiannya menggunakan *cross validation* yang pengujiannya dilakukan 10-fold.

Untuk mendapatkan variable-variabel yang tepat dan menghasilkan nilai akurasi yang terbesar diperlukan pengaturan untuk parameter-parameter genetic optimization.

Skema seleksi yang digunakan adalah roullette wheel. Sedangkan crossover type-nya uniform. Untuk jumlah generasi (number of generation), adjustment dimulai dari 10 – 100, untuk Pop size, adjustment dimulai dari 5 – 50, Untuk P crossover, adjustment dimulai dari 0.1 – 1.0, P mutation dari -1.0 – 1.0.

TABEL I
RENCANA EKSPERIMEN

Num of generation	***Pop Size***	***P Crossover***	***P Mutation***	*Jumlah Atribut*	*Akurasi*
10- 100	5- 50	0.1- 1.0	- 1.0-1.0	?	?

Evaluasi akurasi dilakukan oleh confusion matrix dan curve ROC baik prediksi menggunakan naïve bayes maupun naïve bayes dengan fitur seleksi algoritma genetika. Sedangkan pengujian dilakukan dengan 10-fold validation.

Formulasi akurasi, recall dan precision adalah sebagai berikut :

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{..... (4)}$$

$$\text{Recall} = \frac{tp}{tp+fn} \quad \text{..... (5)}$$

$$\text{NPV} = \frac{tn}{tn+fn} \quad \text{..... (6)}$$

$$\text{PPV} = \frac{tp}{tp+fp} \quad \text{..... (7)}$$

IV. HASIL DAN PEMBAHASAN

Percobaan kedua Naive bayes melakukan training terhadap data-data yang telah dibagi oleh *cross validation.* Setelah dilakukan training dan testing dapat dihitung akurasi dari penerapan algoritma genetika dan naive bayes untuk proses Klasifikasi Abstrak TA. Skema seleksi yang digunakan adalah roulette wheel.

TABEL II
HASIL ANALISA MODEL *NAIVE BAYES* DENGAN *NAIVE BAYES* BERBASIS *GENETIC ALGORITHM*

	naive bayes	**Naive bayes – Genetic Algorithm (GA)**
Recall Farmasi	98%	100%
Recall Kebidanan	88%	96%
Recall Akuntansi	86%	94%
Recall Komputer	88%	100%
Precision Farmasi	92.45%	92.59%
Precision Kebidanan	91.67%	100%
Precision Akuntansi	89.58%	97.92%
Precision Komputer	86.27%	100%
Accuracy	**90.00% +/- 0.00% (mikro:90.00%)**	**97.50% +/- 2.50% (mikro: 97.50%)**

Percobaan dimulai dengan melakukan *adjustment* pada nilai *maximum number of generation*, yaitu dimulai dari 10-100 dengan kelipatan nilai 10, untuk menentukan jumlah generasi yang menghasilkan akurasi paling tinggi. Ketika *maximum number of generation* di-*adjustment*, nilai *pop size, p mutation, dan p crossover* berada pada nilai default, yaitu 5 untuk *pop size*, -1.0 untuk *p mutation*, dan 0.5 untuk *p crossover*. Setelah didapatkan jumlah maksimal generasi yang menghasilkan akurasi paling tinggi, kemudian dilanjutkan dengan melakukan *adjustment* pada nilai *pop size* yang dimulai dari 5-50 dengan kelipatan nilai 5. Nilai pop size yang menghasilkan akurasi paling tinggilah yang akan digunakan pada langkah percobaan selanjutnya. Setelah itu, dilakukan *adjustment* pada *pc* dengan range 0.1-1.0 dengan kelipatan nilai 0.1. Dan yang terakhir dilakukan *adjustment* pada *pm* dengan range -1.0-1.0 dengan kelipatan nilai 0.1.
Berdasarkan dari analisa pengujian antara model *naive bayes* dengan *naive bayes* berbasis *Genetic Algorithm* maka dapat dirangkumkan hasilnya pada tabel 2:

ISSN: 2579-9045
ISBN: 978-602-74355-1-3

V. KESIMPULAN

Dalam penelitian ini dilakukan pengujian model dengan menggunakan *naive bayes* dan *naive bayes* berbasis *Algoritma Genetika* dengan data abstrak. Model yang dihasilkan diuji untuk mendapatkan nilai *accuracy, sensitivity/recall, specifity, PPV, dan NPV* dari setiap algoritma, sehingga didapat pengujian dengan menggunakan *naive bayes* didapat nilai *accuracy* 90.00%, *NPV* 40.00%, *recall farmasi* 98.00%, *recall kebidanan* 88.00%, *recall akuntansi* 86.00%, *recall komputer* 88.00%, *PPV Farmasi* 92.45%, *PPV Kebidanan* 91.67%, *PPV Akuntansi* 89.58%, *PPV Komputer* 86,27%. Sedangkan pengujian dengan menggunakan *naive bayes* berbasis *Algoritma Genetika* didapatkan nilai *accuracy* 97.50%, *NPV* 80.00%, *recall farmasi* 100%, *recall kebidanan* 96.00%, *recall akuntansi* 94.00%, *recall komputer* 100%, *PPV Farmasi* 92.59%, *PPV Kebidanan* 100%, *PPV Akuntansi* 97.92%, *PPV Komputer* 100%. Maka dapat disimpulkan pengujian model Klasifikasi Abstrak Tugas Akhir dengan menggunakan *naive bayes* dan *naive bayes* berbasis *Algoritma Genetika* didapat bahwa pengujian *naive bayes* berbasis *Algoritma Genetika* lebih baik daripada *naive bayes* sendiri.

Dengan demikian dari hasil pengujian model diatas dapat disimpulkan bahwa *naive bayes* berbasis *Algorritma Genetika* memberikan pemecahan untuk permasalahan Klasifikasi Abstrak Tugas Akhir lebih akurat.

DAFTAR PUSTAKA

[1] M. D. R. R. M. Dr.Ir.Bambang Dwiloka, Teknik Menulis Karya Ilmiah, Jakarta: Bhineka Cipta, 2012.
[2] W. Chang., Metodologi Penulisan Ilmiah, Jakarta: Erlangga, 2014.
[3] Han, J., & Kamber, M. *Data Mining Concepts and Techniques*. San Francisco: Diane Cerra, 2007.
[4] B. Santoso, Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis, Yogyakarta, 2007.
[5] M. Brammer, *Principles of Data Mining*. London: Spinger, 2007.
[6] Vivek Narayan, "Fast And Accurate Sentiment Classification Using an Enhanced Naive Bayes Model," *Spinger Verlag Berlin*, pp. 194-201, 2013.
[7] K. P. K. a. J. P. Vishwanath Bijalwan, "KNN based Machine Learning Approach for Text and Document Mining," *International Journal of Database Theory and Application*, vol. 7 no. 1, pp. 61-70, 2014.
[8] M. M. M. Ms S. Vijayarani, "Comparative Analysis of Bayes and Lazy Classification Algorithms," *International Journal of Advanced Research in Computer and Communication Engineering* , vol. 2, no. 8, 2013.
[9] Feldman. R. &. Sanger. J., The Text Mining Handbook Advance Approaches in Analyzing Unstructured Data, Cambridge University Press, 2007
[10] Bakti, V.K. and Indriyatno, J., 2017. Klasterisasi Dokumen Tugas Akhir Menggunakan K-Means Clustering, Sebagai Analisa Penerapan Sistem Temu Kembali. *KOPERTIP: Jurnal Ilmiah Manajemen Informatika dan Komputer*, *1*(1), pp.31-34.